Preparing for Google Cloud Professional Data Engineer, path for data engineering for DBAs.

# About Myself

My Name is Namit Sehgal, 17+ Years of hands-on experience as extremely passionate DevOps DBA / Database Solution Architect / Data Engineering in both, RDBMs and NoSQL database world. Responsible for full databases life cycle starting form finding optimum database specific appliances, performing server setups, database designs, implementations of various project with focus on optimization, protection, database integrity and disaster recovery, releases using CI/CD pipeline tools. Trained and navigated other DBAs.

And...

Changed my role from DBA to a Application DBA 8 years back.

And..

I started my cloud journey 4 years back and now doing project management, Solution Architect, Devops, APPs integration on Cloud and now stabilizing my career or looking ahead myself to be a Data Engineer.’

And still want to have Data as core component in my career.

And

I love to train my Daughter in Maths and now spending time with my new born.

And

Love eating different food, favourite is Thai.

And

My wife may not agree but I was the one who helped her in the cooking, recently not taking much part in cooking as I am helper is doing that now. Maybe as Data Engineer I will be doing data prepare..

And doing Yoga, running to stay healthy.

# Journey for a DBA in Cloud

With the evolution of cloud, I believe that we do not have any roles for infrastructure DBAs and they must see where the tide is moving and must get trained to go along with that tide.

I always thought that a good DBA is the one who can look at a database not only in terms of infra perspective but also application specific. Some myths I always believe that how to fix issues with database without switching on or off the power, how best you can optimize the database with existing hardware, managing within the exiting costs. Believe it or not, 90 percent of issues get resolved with DBAs as they can only give proper recommendations even to developers. So issue start with balining database and ends with taking recommendation.

So ***future of DBA*** is how they can migrate existing workloads to cloud and once we are on cloud, are we going to lose job again? No, life keeps on coming up with new evaluation and give human beings opportunities, atleast think we are better than System Admins, Storage Admins, and ofocurse those managers who just get alive during incidents to know the status of the issue. Anyways they also have to be technical now to survive….Singal to various IT companies who provide IT services to get rid of such managers or ask them to retrain.

What next for DBAs?

During my cloud journey I found that migrating databases and doing POC with PaaS databases provided by various vendors may end soon but there is something which all of data savvy people can do is to look at a role called *Data Engineer. It is vast and this role has a lot of space for us to accommodate, We can use TSQL to write AI models, isn’t that what we know for ages?*

Enough of DBAs, I think this article is how we can do GCP Data Engineer.

# Why GCP Data Engineer for a start?

I am working extensively in Azure and I think my Microsoft colleagues will say that I know something in Azure now..

Cosmos, a lot of cost, buying products and use in under MS umbrella…but tahts what all companies do.

Azure Datawarehouse or new name Synapse, still evaltuing. I do not see any difference in traditional SQL Server.

AWS, first to come in the market and came up with DynampDB, Aurora etc…

Again, coming into first in the market is an advantage but staying at top is challenging.

Oracle, I love it, no other database as powerful as Oracle. But did they made a mistake hiring many VPs who never work as a core DBA.I think company strategy of promoting or hiring wrong people made Oracle cloud to sit behind. Just an example. With so much of money, they took ages to open a datacentre in Singapore. No one waits Oracle.

I still say, I love Oracle.

Google was the one who wrote paper on Big data, they are the ones came up with Big Table and later with Big Query. Intelligent people took their paper and created other NoSQL databases. So why not start my journey with the creator of handling problems of Big Data.

Some question we always have, how to scale my RDBMS horizontally?

How can someone manage my databases without indexes. Update stats?

Isn’t that column store database is much powerful?

How can I secure my data with so many hackers around?

In my current job, we got second cloud provider as Google and knowing the names of Big Query, Cloud Spanner, Big Table made me think that I need to know more about this.

I spoke to my Google colleagues and they told me there is only one course which I can do related to data.

* You might have seen this:-

Certificate Exam Guide:-

<https://cloud.google.com/certification/guides/data-engineer/>

* With no experience in ML and AI, I was worried if I can do this but did I have any other choice? You also do not have… earlier the better.
* I liked this as well:-

Notes for GCP Data Engineer

<https://sathishvj.medium.com/notes-from-my-google-cloud-professional-data-engineer-exam-530d11966aa0>

Following courses I took:-

Share subscriptions with your friends so you need to pay for every cloud training provider annual cost. Bad thing is you cannot buy courses, you have to take full subscription.

* https://app.linuxacademy.com/

Course: Google Certified Professional Data Engineer

Course: Google Cloud Certified Professional Data Engineer (LA)

* <https://app.pluralsight.com/>

Google Cloud Certified Professional Data Engineer

Check your Skill Level in Data Engineering on Google Cloud and then do courses once your evaluation is done with at pluralsight.

* Do practice here

<https://www.examtopics.com/exams/google/professional-data-engineer/view/4/>

Issue here is answers are not verified by Google but you can click on feedback and see what people are talking about.

* <https://www.udemy.com/>

Google Cloud Professional Data Engineer: Get Certified 2020 by Build scalable

* Exam will have 10 to 15 percent question on Google Big Query
* Exam will have 10 to 15 questions on AI and ML.
* Many question which will start with keyword ETL, so make sure you understand full life cycle of ETL in Google and overall.
* You must make up your mind that questions can be general fundamental questions, not only related to Google services. For example, you can exp[ect in Exam:-

1. Selling a house, price goes higher with latitude and longitude of the house, so study features Engineering.
2. You are using Google Home and you give an instruction to switch off your fan, which Model in GCP is the best? All answers seem correct…Choose the best suitable one. Don’t worry about result.

* In all your practice exams in above courses, make sure you score atleadt 60 percent in first attempt. Answers which are wrong, understand Why?
* Next attempt target to have above 90 percent, but do not do immediately as you will remember the answers, switch to different course once you attempt first exam.=
* Please find attached my Notes, I am sure you can add your own notes once you do the course but read it before 1 hour yu go for Exam.
* During the Exam, giving outside, you will be asked to wear gloves and can not remove Mask. Do not panic.
* Mark all answers which you have any doubts for review. In my case, I marked 35 for review as I was sure for only 15 answers.
* No idea how much is passed percentage, So target to score 80 percent in your pratice exams.
* 2 hours is not enough, so end of the Exam, give Google feedback.
* Use bigger font size; you can adjust in the beginning of exam. Bigger make it faster to read, better for my eyes.
* Do not rush for Exam and take atleadt 2 to 3 months for preparation, It is not AWS or Azure exam which you can do in 10 days preparation and many people are going for these as they are easier to crack. Try something which is challenging and which is going to provide you data solutions for various problems.
* Learning Al and ML fundamentals is not easy. Watch youtube for things like evalution of neural networks.
* Watch videos at youtube.com for Google Cloud services for example Big Query, Spanner etc.

What next?

In your organization we you are using Azure, prepare for 200 and 201 and do similar for AWS.

Find a job where you can do 70 to 80 percent of your job in data preparation and then it will make easier for you to jump into Data Sceince.

|  |  |
| --- | --- |
| **Item** | **Comment** |
| Sibnets extend across regions in GCP. | Data Enginner to know design infrastructure in GCP across regions. |
|  | JSON instacnes can be in same subnets but across zones. |
|  | a subnet spans zones, enabling VMs with adjacent IPs to exist in separate zones, making design for availability easier to accomplish since the VMs can share tagged firewall rules." |
| Certificate Exam Guide | <https://cloud.google.com/certification/guides/data-engineer/> |
| Cloud Storage, Cloud Datastore. Cloud firestore and Cloud Spanner, Bigquery, Cloud dataflow are serverless. |  |
| CloudSQL and Cloud Bigtable, Cloud Dataproc are managed services. | Managed servcies still has some overhead but minimizes it comapring to on-premise. |
| CloudSQL for consitent storage but not for large amoint of unstructured data for example video files. |  |
| Cloud data proc, data abstraction with sparc is resinet distiributed dataset(RDD), processing abstraction is directed acyclic grpah (DHE) |  |
| Bigquery absractions are tables and queries. |  |
| Dataflow absractions are pcollection and pipeline. |  |
| Cloud pub sub can hold a message for 7 days. |  |
| N4 standard VMs have more IOPs / cost then N1 instacnes. |  |
| Cloud Storage is an Object stored in Bucket. |  |
| Data in Cloud Datastore is a property, contained in an Entity and is a KIND category. |  |
| Data in Cloud SQL is storage in rows, colimns and in tables. |  |
| Data in Cloud spanner consists of Values stored in columns and rows in a table. |  |
| Avaro serializes data to Binary format. |  |
| Batch data is bounded data(usually a file), stream data is unbounded data (dynamically generated). |  |
| pCollections and RDD are not identical. |  |
| Data in Tensorflow is represented in tensors. |  |
| Tensor Rank 0(scaler),Tensor Rank 1(vector),Tensor Rank 2(matrix),Tensor Rank 3(cube) |  |
| Cloud Dataproc (HBASE >> Cloud Bigtable, HDFS >> Cloud Storage), HA (3 masters), Standard(1 master only),preempitable VMs. |  |
| Spark-part of processing in memeory. |  |
| Cloud Dataflow(Pardo- Allow for parallel processing, Map- 1:1 relationship with input and output in Puthon, Flatmap-non 1:1 relationship with input and output(generator with Python), Combine (aggregate value), Groupby- Shuffle, ,apply(Pardo)- Java for Map and FlatMap, Groupbykey- Explcitly Shuffle. Side Inputs- Parallel Paths of execution for making different transformations on the same source data. |  |
| Dataflow - Developers create pipelines>> stores templates in cloud storage >> consumers runs for Production. |  |
| Dataflow pipelines are ofetn recoignized as Map and Reduce sequences. |  |
| Transfer data(gsutils, Tansfer Appliance- transfer on-premise, Storage Transfer Services- Other Cloud privdrs) |  |
| Cloud Storage can be used for Traffic Esstimation. |  |
| A bigquery dataset needs to be multi-regional after configured as regional. Use backup to nearby cloud storage and restore in another region. |  |
| Notes for GCP Data Engineer | <https://sathishvj.medium.com/notes-from-my-google-cloud-professional-data-engineer-exam-530d11966aa0> |
| 3 modes for Natural Language API( Sentimental analysis, Entities and Syntax), can be used identifying all locations mentioned in document. |  |
| Superwise learning - Questions on Labels, Superwise Machine Learning - Classfication and Regression, |  |
| Consider data where it is, do not use ETL. |  |
| How is the CLUSTER\_SPEC formatted in Cloud ML Engine? | By cluster, task, and job |
| What is the counterpart to REPEATED (Legacy SQL) in Standard SQL? | ARRAY |
| What are the two request formats for Speech-to-Text responses? | REST and gRPC |
| In order to create a pipeline’s initial PCollection, what transform must you apply to your pipeline object and why? | A root transform in order to read pipeline data from an external data source or create pipeline data from an in-memory source. |
| When should you use a `languageHints` object for OCR in Cloud Vision API? | When the language to be recognized is absolutely known. |
| Datastore - Entities of the same kind can have properties of the same name and with different types |  |
| Which of the following best describes Pub/Sub delivery guarantees? | At least once |
| Which of the following best describes Datalab? | GCP service that allows users to run interactive Python notebooks on the cloud; runs on GCE VM instance and based on Jupyter (formerly iPython) |
| What is another name for Batch prediction? | Offline prediction |
| What is the purpose of a window in Cloud Pub/Sub? | To split the infinite stream into finite buckets. |
| Which of the following describes Cloud Spannerâ€™s data model? | All tables must include key columns; data from parent and child tables are interleaved, and can be moved around by Cloud Spanner in independent units (called splits) to improve performance |
| Default services in GCP for ML. | Cloud Vision API Cloud Speeach to text API Cloud Video Intellignece API Cloud Natural Language API Cloud Translation API Cloud Dialogflow is used for creating chatbox. |
| To run a local training job using the Google Cloud SDK, what command would you run? | gcloud ml-engine local train |
| Wide neural networks are better for memorization. |  |
| How do you provide GCP authentication credentials when making request to Cloud MP APIs from the command line with curl? | Configure authentication with gcloud, then use 'print-access-token' to include an authorization bearer in the HTTP request. |
| How should you interpret the documentSentiment response from the Natural Language API? | score' represents an overall emotional leaning of a test from -1.0 (negative) to 1.0 (positive), and 'magnitude' indicates the overall strength of emotion. 'magnitude' is not normalized so longer text blocks may have greater magnitudes. |
| How would you begin to design a Dialogflow agent that can answer requests for weather forecasts? | Create an Intent for requesting a weather forecast, specify training phrases based on real-world speech patterns that would ask this question, and define the structured parameters that can be used to create a response. |
| If you make a request to the documents:analyzeEntities endpoint of the Natural Language API, what sort of information can you expect in the response? | An array of Entity objects that will each contain metadata including a Wikipedia URL and Knowledge Graph MID, if available |
| Which AutoML Vision feature would you use to detect multiple objects in an image? | Object localization |
| You have a large number of images that you wish to process through a custom AutoML Vision model. Time is not a factor, but cost is. Which approach should you take? | Make an asynchronous prediction request for the entire batch of images using the batchPredict method. |
| You wish to build an AutoML Natural Language model for classifying some documents with user-defined labels. How can you ensure you are providing quality training data for the model? | Ensure you provide at minimum 10 training documents per label, but ideally 100 times more documents for the most common label than for the least common label. |
| Hive is a data warehouse application that uses SQL queries, which is exactly what BigQuery performs. |  |
| Dataflow IAM | You can control access at the publisher level. |
| How can you connect to the web interface of a Dataproc cluster? (Choose two) | SOCKS proxy, Allow the necessary web ports access via firewall rules, and limit access to your network. |
| Which operation can be used to invoke a user-specified function of each element of an input PCollection? | ParDo |
| What is the purpose of a trigger in Cloud Dataflow? | Triggers determine when to emit output data, and behave differently for bounded and unbounded data. Triggers determine when to emit aggregated results as data arrives. For bounded data, results are emitted after all of the input has been processed. For unbounded data, results are emitted when the watermark passes the end of the window, indicating that the system believes all input data for that window has been processed. |
| Which transformation can be used to process collections of key/value pairs, in a similar fashion to the shuffle phase of a map/shuffle/reduce-style algorithm? | GroupByKey is a Beam transform for processing collections of key/value pairs. It’s a parallel reduction operation, analogous to the Shuffle phase of a Map/Shuffle/Reduce-style algorithm. CoGroupByKey performs a relational join of two or more key/value PCollections that have the same key type. Combine simply combines elements, and Partition splits elements into smaller collections. |
| Cloud Dataproc clusters can be provisioned with a custom image that includes a user's pre-installed packages. You could alternatively use initialization actions to install the additional components, but this would be less efficient and incur more running time for ephemeral clusters. | Create custom Dataproc image that fulfils the customer requirements and use it to deploy a Dataproc cluster. |
| Preemptible workers in a Dataproc cluster cannot store HDFS data. |  |
| if you have 2 replicating clusters in your Bigtable instance, how can you ensure that your application will be guaranteed strong consistency for its transactions? | Use an application profile that specifies single-cluster routing. Strong consistency can only be achieved using single-cluster routing. Eventual consistency is normally quick but can take several minutes depending on the distance between clusters. If your application requires strong consistency, refactoring is unlikely to be an option without a complete redesign. |
| What is the maximum number of clusters per Bigtable instance? | A Bigtable instance can contain up to 4 clusters. Compute Engine is a red herring, it has nothing to do with Cloud Bigtable. |
| What is the minimum retention duration for a message in Pub/Sub? | Message retention duration can be configured between 10 minutes and 7 days. |
| A push Subscription requires what as its endpoint? | A push Subscription endpoint must accept an HTTPS POST with a valid SSL certificate. It must also be configured to use an authentication header. |
| You want to use Pub/Sub to distribute jobs to a group of Compute Engine VMs, which should each take the next job from the queue. How should you configure Pub/Sub? | Create a Topic for jobs. Create a Subscription for this Topic that can be shared by every Compute Engine VM in the group. |
| You are writing a streaming Cloud Dataflow pipeline that transforms user activity updates before writing them to a time-series database. While continually transforming each element as it arrives, you also need to depend on some additional data at run-time to create the transformation. How can you achieve this? | Side inputs are useful if your ParDo needs to inject additional data when processing each element in the input PCollection, but the additional data needs to be determined at runtime (and not hard-coded). A combine would not achieve the same outcome, and using an external shell script is unnecessary and inefficient. |
| Cloud Pub/Sub is an ideal managed replacement for Kafka. It is globally available so topics do not need to be created on the basis of regio |  |
| Large numbers of cells in a row can cause poor performance in Cloud Bigtable. When the ***data itself is so small, as in this scenario, it would be more efficient to simply retrieve all of the metrics from a single cell, and use delimiters inside the cell to separate the data.*** Row versioning would compound the problem by creating the most new entries along the least efficient dimension of the table, and HDD disks will always slow things down |  |
| Order tables appropriately in the query, with the larger table on the left side of the JOIN and the smaller table of the right side of the JOIN. |  |
| Which of these is NOT a type of trigger that applies to Dataflow? | Element size in bytes. |
| Using the --no-address flag will prevent public IPs from being assigned to nodes in a Cloud Dataproc cluster. However, Private Google Access is still required for the subnet to access certain GCP APIs. |  |
| You must set Owner credentials to use the enable cache option in BigQuery. It is also a Google best practice to use the enable cache option when the business scenario calls for using prefetch caching.Cache auto-expires after 12 hours. 24-hour cache is not a valid option. Prefetch cache is only for data sources that use the Owner's credentials (not the Viewer's credentials) |  |
| The correct answer is to attach TPUs to Compute Engine VMs to accelerate Tensorflow model training. Training Tensorflow models on very large data sets using CPUs is not as cost efficient as using TPUs so scaling the VMs, either horizontally or vertically, is not as cost effective as using TPUs. Simply deploying the training workload to Kubernetes without accelerators will not significantly improve performance the way TPUs can. |  |
| The correct answer is the statement that references \_PARTITIONTIME, which is the pseudo-column created by BigQuery for tables partitioned on ingestion. PARTIONTIME, \_INGESTIONTIME, and INGESTIONTIME are not the correct name of the pseudo column. |  |
| The correct answer is to migrate the data to Bigtable and use Bigtable's HBase API. Cloud Storage, Cloud Datastore, and Bigquery do not have an HBase API. |  |
| The correct answer is ***stochastic gradient descent because it updates after each instance is analyzed. Batch gradient descent is incorrect because it updates after processing the entire training set. Mini-batch is incorrect because it updates after multiple instances. (The number of instances is less than the number of instances in the training set.)***  Max-batch gradient descent is not the name of a form of gradient descent. |  |
| The correct answer is ***design rows to store the set of measurements from one sensor at one point in time.*** These are known as narrow tables. Designing rows to store measurements from one sensor over an hour is a wide-table pattern but not recommended for time series data. Designing rows to store measurements from all sensors at a single time point would be a wide table pattern and require waiting for all or most of the sensor values to arrive before writing the data. ***Designing rows up to the maximum size of 100 MB per row would also lead to a wide table pattern.*** |  |
| You are migrating a data warehouse to BigQuery and want to optimize the data types using in BigQuery. You have many columns in the existing data warehouse that store absolute point in time values. They are implemented using 8-byte integers in the existing data warehouse. What data type would you use in BigQuery? | The correct answer is to use a timestamp to represent an absolute point in time. Long integer is not an available data type, although INT64 and NUMERIC could be used to represent long integers, i.e. 8-byte integers. Datetime is used for data and time independent of timezone. Time is used to represent clock time. |
| ***HyperLogLog, which is a data structure that provides approximate distinct item counts to set with low latency.*** Sorted Sets will not reduce the time to found distinct items. There is no Stochastic Set data type in Redis. List will not provide approximate counts of distinct items. |  |
| The correct answer is 'gcloud dataproc jobs submit spark --driver-log-levels'. 'gcloud dataproc submit jobs spark --driver-log-levels=debug' is incorrect because the order of 'jobs' and 'submit' are reversed. The other options are incorrect, there is no --enable-debug option. |  |
| A Python ETL process is loading a data warehouse is not meeting ingestion SLAs. The service that performs the ingestion and initial processing cannot keep up with incoming data at peak times. The peak times do not last longer than one minute and occur at most once per day but data is sometimes lost during those times. You need to ensure data is not lost to the ingestion process. What would you try first to prevent data loss? | The correct answer is to ingest data into a Cloud Pub/Sub topic using a pull processing mode. Cloud Pub/Sub will scale as needed and will not lose data. By having the consuming service pull messages when it is able to, data is cached until it can be processed. Rewriting an ETL process in another language could be difficult and time consuming and should not be tried before less drastic changes are evaluated. A push subscription is not appropriate since the consuming service presumably would not be able to process peak loads fast enough to keep up. Cloud Dataflow is used for stream and batch process and not scalable ingestion pipelines. |
| You support an ETL process on-premises and need to migrate it to a virtual machine running in Google Cloud. The process sometimes fails without warning. You do not have time to diagnose and correct the problem before migrating. What can you do to discover failure as soon as possible? | The correct answer is to create a Cloud Monitor uptime check and if the uptime check fails send a notification to you. Creating an application on App Engine could work but it requires an additional service to maintain and would incur additional costs. Creating an alert on CPU utilization falling below 5% is incorrect because CPU utilization dropping below 5% does not mean the process failed. There is no way to create an alert on Cloud Logging not receiving data from a service. |
| You have been asked to help diagnose a deep learning neural network that has been trained with a large dataset over hundreds of epochs but the accuracy, precision, and recall are below the levels required on both training and test data sets. You start by reviewing the features and see all the features on numeric. Some are on the scale of 0 to 1, some are on the scale of 0 to 100, and several are on the scale of 0 to 10,000. What feature engineering technique would you use and why? | The correct answer is ***normalization, to map all features to the same 0 to 1 scale. Regularization is a technique used to reduce the amount of information captured by a model to prevent overfitting. The model is not overfitting because it performs poorly on training data. Regularization does not map values to a 0 to 1 scale. Backpropagation is used to calculate the gradient in neural network learning, not for reducing the amount of information captured in a model.*** |
| A business intelligence analyst is running many BigQuery queries that are scanning large amounts of data, which leads to higher BigQuery costs. What would you recommend the analyst do to better understand the cost of queries before executing them? | The correct answer is to use the bq query command with the SQL statement and the --dry-run option to return an estimate of the amount of data scanned. There is no --estimate option in the bq command. The --max-rows-per-request sets the maximum number of rows to return per read. There is no gcloud bigquery command. |
| When testing a regression model, you notice that small changes in a few features can lead to large differences in the output. This is an example of what kind of problem? | ***The correct answer is this is an example of high variance. Low variance is desired and not a problem. High bias occurs when relationships are missed. Low bias is desired and not a problem.*** |
| A number of machine learning models used by your company are producing questionable results, particularly with some demographic groups. You suspect there may be an unfairness bias in these models. Which of the following could you use to assess the possibility of unfairness and bias? | The correct answer is ***classification parity which measures the predictive performance across groups. The more equal the classification parity measure, the less unfair or biased the model is. Anti-classification is a method of avoiding bias but not assessing bias. Regularization is a method used to reduce the risk of overfitting. Normalization is a feature engineering technique.*** |
| Pub / Sub -Seek is used to take care of subscriber upgrades, messages are retained and subscriber can rewind time stamp to get those messages. -Snapshot can be used to replay messages. -Message data should be less than 10 MB |  |
| A neural network with (generally) at least 3 hidden layers |  |
| Keras is a high-level deep-learning Python library that includes support for TensorFlow functionality. |  |
| Example (training) data uses its labels to train the machine learning model, while test data uses its labels to validate the model’s accuracy. |  |
| hyperparameters are the variables that govern the training process itself. They are configuration variables, not directly related to the training data. For example  Hidden layers, Learning rate |  |
| A Tensor is a multi-dimensional array. In TensorFlow, tf.Tensor objects have a data type and a shape and can be stored in accelerator memory (ie. GPU memory). |  |
| L1 regularization - A type of regularization that penalizes weights in proportion to the sum of the absolute values of the weights. In models relying on ***sparse features***, L1 regularization helps drive the weights of irrelevant or barely relevant features to exactly 0, which removes those features from the model. |  |
| L2 regularization- A type of regularization that penalizes weights in proportion to the sum of the squares of the weights. L2 regularization helps drive outlier w**eights (those with high positive or low negative values)** closer to 0 but not quite to 0. (Contrast with L1 regularization.) L2 regularization always improves generalization in linear models. |  |
| logistic regression - A classification model that uses a sigmoid function to convert a linear model's raw prediction () into a value between 0 and 1. |  |
| discriminative model-  a model that predicts whether an email is spam from features and weights is a discriminative model. The vast majority of supervised learning models, including classification and logical regression models, are discriminative models |  |
| dynamic model - A model that is trained online in a continuously updating fashion. That is, data is continuously entering the model. |  |
| Features are inputs and Labels are categories or entities which we are going to predtict with these fatures. Mean square error used for regrssion problem. Accuracy, percicsion and recall used for classifcation problem. Hyperparameter is what we tune, theya re not parameters like weights. Learning rate is hyperparameter. epochs are training time. ***Bias is error in missing relationship between feature and labels. Varinece is senstivity in small flucatations in the training dat***a. | <https://developers.google.com/machine-learning/glossary> |
| reinforcement learning  A family of algorithms that learn an optimal policy, whose goal is to maximize return when interacting with an environment. For example, the ultimate reward of most games is victory. Reinforcement learning systems can become expert at playing complex games by evaluating sequences of previous game moves that ultimately led to wins and sequences that ultimately led to losses. |  |
| supervised Learning Learn from examples m Goal is to predict category or values  - Classifying tumours from medical images (Classification Model)  - Predicting housing prices (Linear Regression)  - Identifying fradulent credit card trasactions |  |
| overfitting- Creating a model that matches the training data so closely that the model fails to make correct predictions on new data. underfitting- Producing a model with poor predictive ability because the model hasn't captured the complexity of the training data. |  |
| regularization The penalty on a model's complexity. Regularization helps prevent overfitting. Different kinds of regularization include: L1 regularization L2 regularization dropout regularization early stopping (this is not a formal regularization method, but can effectively limit overfitting) |  |
| collaborative filtering Making predictions about the interests of one user based on the interests of many other users. Collaborative filtering is often used in recommendation systems. |  |
| dimension reduction Decreasing the number of dimensions used to represent a particular feature in a feature vector, typically by converting to an embedding. |  |
| matrix factorization #recsystems In math, a mechanism for finding the matrices whose dot product approximates a target matrix.  In recommendation systems, the target matrix often holds users' ratings on items. For example, the target matrix for a movie recommendation system |  |
| Dataflow Windows"- Event TIme is time when data elemant occurs. pcollection, Pardo (filter data for example extract some columns) Time Based Windows  Watermark trcks how far behind the system in processing data from the evnet time. Fixed, Sliding (running Avg) and session based windows, Single global window (default) . Triggers when to give off (Event, processing,data, composite), what to do with late arrving data. Watermark ***What are side windows? Side Inputs- Parallel Paths of execution for making different transformations on the same source data.***   ***Try catch block to collect error data.*** Update existing pipeline which takes care of infly data. Regional endpoint for dataflow is required. FlexRS (Advanced schduling, Cloud dataflow shuffle serice, Premptible VMs) |  |
| Cloud Firestore and Datastore -Realtime DB with mobile SDKs - Collection contain documents but document can also contain sub collections. - References can be used instead of su b collections to efer data in another document. - Automatic single field indexes -Index exemption - composite indexes can be created - User profiles store foe example gaming Datasrore(table - Kind, Row - Entity, Column - property, Primary key = key) Entity can have nested entities can not change region to multi reional once defined. |  |
| Cloud Spanner (CAP) - Managed SQL - Horizontally scalable - High avaiaiable  - 3 read write replicas |  |
| Unsupervised Learning (draw inference from data), previously undetected patterns. No Labels  -Clustering  -Anomoly detection - Timeseries data - Principal componete analysis |  |
| Transfer learning is trained model on short amount of data |  |
| Cloud Speech-to-text API - Sync, Asyn, Streaming - REST and gRPC while streaming is only with gRPC |  |
| Dialogflow - Intant classfication - Training Phase >> Model >> extract patameter - Returns confidence and transcript |  |
| Cloud Text-to-speech API - SSML |  |
| Cloud AutoML Natural Language \_ Translate Language - Extract meaning of Language - Sentimental analysis, entity analysis, Syntax, Entity sentiment, Content classfication - Score is from -1 to 1, Magnitude s not normalized ranges from 0 to infinity  - Entities are nounus , identifies wikipedia url if it is there. - Syntax ( sentecnes and tokens) |  |
| Cloud Video Intelligence API  - Detect Labels   - Shot change detection  - Detect explicit content - Speech transcript - Object Detection - Text detection |  |
| Cloud Vision API - Sync or Async mode  - OCR Text detection and coeumtn text detection (extract text from images) - Face detection (emotional state or headwear) - Celebrity detection  - Label detection with a Score  - Landmarh detection  - Logo detection  - Adult ot violenet contect (safe serack detection 5 categories) - Web and Page detection |  |
| BigQuery can export Avro data natively to Cloud Storage. **BigQuery Transfer Service is for moving BigQuery data to Google SaaS applications (AdWords, DoubleClick, etc.).** You will want to perform a normal export of data, which works with Avro formatted data. BigQuery can export Avro data natively to Cloud Storage. BigQuery can export Avro data natively to Cloud Storage. |  |
| Cloud Spanner automatically creates an index on the primary key. For efficient searches of both authors and genres, it would make sense to use authors as a primary key, and add a secondary index for the genre column. Creating the index when the data is loaded is the most efficient way to achieve this |  |
| By default, preemptible node disk sizes are limited to 100GB or the size of the non-preemptible node disk sizes, whichever is smaller. However, you can override the default preemptible disk size to any requested size. Since the majority of our cluster is using preemptible nodes, the size of the disk used for caching operations will see a noticeable performance improvement using a larger disk. Also, SSD's will perform better than HDD. CSV format would perform worse than parquet. Increasing the cluster size may also improve performance, but at a greater additional cost than these other steps |  |
| Bigtable Syncing service option does not exist. Cloud Storage works as an in-between for the two services. However, **you can directly interact between Dataproc and Bigquery using the Bigquery connector.** |  |
| CoGroupByKey performs a relational join of two or more key/value PCollections that have the same key type; "common key" is the magic clue in this question. |  |
| In a Dataflow processing pipeline, which concept describes timestamps attached to incoming messages? | Watermark describes the event time, which is what a timestamp designates. ParDo is part of the Apache Beam workflow but does not describe timestamps. PCollection is part of the Apache Beam workflow but does not describe timestamps. Triggers are what determine when a window's contents should be submitted. |
| Your organization will be deploying a new fleet of IoT devices, and writes to your Bigtable instance are expected to peak at 50,000 queries per second. You have optimized your row key design and need to design a cluster that can meet this demand. What do you do? | An optimized ***Bigtable*** instance with a well-designed row key schema can theoretically support up to ***10,000 write queries per second per node, so 5 nodes are required.*** |
| Since **you cannot change the disk type on an existing Bigtable instance,** you will need to export/import your Bigtable data into a new instance with the different storage type. You will need to export to Cloud Storage then back to Bigtable again. |  |
| Organization level roles are way too broad for this scenario. It is ***best practice to use pre-created roles over custom roles*** with associated policies when they match your requirements, which they do in this scenario, and the principle of least privilege favors using pre-defined roles for granular access. |  |
| ***L2 regularization is more relevant when all features have relatively equal weights/influence,*** which is not the case here. Hyperparameters deal with learning rate, which is not relevant for this question. **L1 regularization is able to reduce the weights of less important features to zero or near zero.** Epochs are related to training time, which is not relevant here |  |
| Deep and wide models are ideal for a recommendation application. |  |
| Reinforcement learning uses reward systems to complete a task. Predicting the location would be an example of using unsupervised learning to find patterns. ***Unsupervised learning does not use labels but does look for patterns (or clustering) of data in order to make predictions based on the patterns it learns.*** Supervised learning takes labeled training data to predict future results on new test data. Classification models apply categorized variables (i.e. "fraudulent", "not fraudulent"). |  |
| Your organization needs to develop their machine learning model to control topology definitions. There are a large number of possible configurations to achieve the best results. What components of their machine learning model would they adjust to account for increased complexity? | Learning rate is a hyperparameter, not related to adjusting to training data. An Epoch is a pass through the training dataset, not related to complexity. ***Adding additional neurons allows combining more input values, and hidden layers allow for passing outputs to another layer of neurons for more complex calculations.*** |
| You want to train your machine learning model on AI Platform while saving costs. Which scaling tier would you choose? | BASIC tier uses a single instance and is the lowest cost. CUSTOM can use different configurations, but BASIC is still cheaper with a single instance. STANDARD\_1 and PREMIUM\_1 all use multiple worker and parameter servers are are more expensive |
| Which of the following best describes Datalab? | GCP service that allows users to run interactive Python notebooks on the cloud; runs on GCE VM instance and based on Jupyter (formerly iPython) |
| What is another name for Batch prediction? | Offline prediction |
| What is the purpose of a window in Cloud Pub/Sub? | To split the infinite stream into finite buckets. Incorrect |
| Why does the performance of BigTable often improve over time? | BigTable figures out patterns in queries and evenly redistributes data to better parallelize reads and writes |
| Which of the following describes Cloud Spannerâ€™s data model? | All tables must include key columns; data from parent and child tables are interleaved, and can be moved around by Cloud Spanner in independent units (called splits) to improve performance |
| 16. When transferring Hadoop jobs to GCP. After, you update your job to point to persistent data stored in Cloud Storage. When creating a Cloud Dataproc cluster on which to run your job, what must you do next? | Submit your job to the cluster for processing. |
| Cloud Spanner offers great transaction support, even stronger than ACID. How best to avoid paying the performance penalty for this when not required? | Use ***single read calls for one-off reads,*** and read-only transactions wherever possible |
| Dataprepr uses dataflow in background. |  |
| How much data can you store in a PCollection? | Virtually Limitless |